Fast Adapting without Forgetting for Face Recognition

Hao Liu, Xiangyu Zhu, Zhen Lei, *Senior Member, IEEE,* Dong Cao, Stan Z. Li, *Fellow, IEEE*

***Abstract*—Although face recognition has made dramatic improvements in recent years, there are still many challenges in real-world applications such as face recognition for the elderly and children, for the surveillance scenes and for Near infrared vs. Visible light (NIR-VIS) heterogeneous scene, etc. Due to the existence of these challenges, there are usually domain gaps between training (source domain) and test (target domain). A common way to improve the performance on the target domain is fine-tuning the base model trained on source domain using target data. However, it will severely degrade performance on the source domain. Another way which jointly trains models using both source and target data, suffers from the heavy computations and large data storage, especially when we continue to encounter new domains. In response to these problems, we introduce a new challenging task: Single Exemplar Domain Incremental Learning (SE-DIL), which utilizes the target domain data and just one exemplar per identity from source domain data to quickly improve the performance on the target domain while keeping the performance on the source domain. To deal with SE-DIL, we propose our Fast Adapting without Forgetting (FAwF) method with three components: margin-based exemplar selection, prototype-based class extension and hard&soft knowledge distillation. Through FAwF, we can well maintain the source domain performance with only one sample per source domain class, greatly reducing the fine-tuning time-cost and data storage. Besides, we collected a large-scale children face dataset KidsFace with $12K$ identities for studying the SE-DIL in face recognition. Extensive analysis and experiments on our KidsFace-Test protocol and other challenging face test sets show that our method performs better than the state-of-the-art methods on both target and source domain.**

***Index Terms*—Single Exemplar, Domain Incremental Learning, Fast Adapting without Forgetting, Face recognition.**

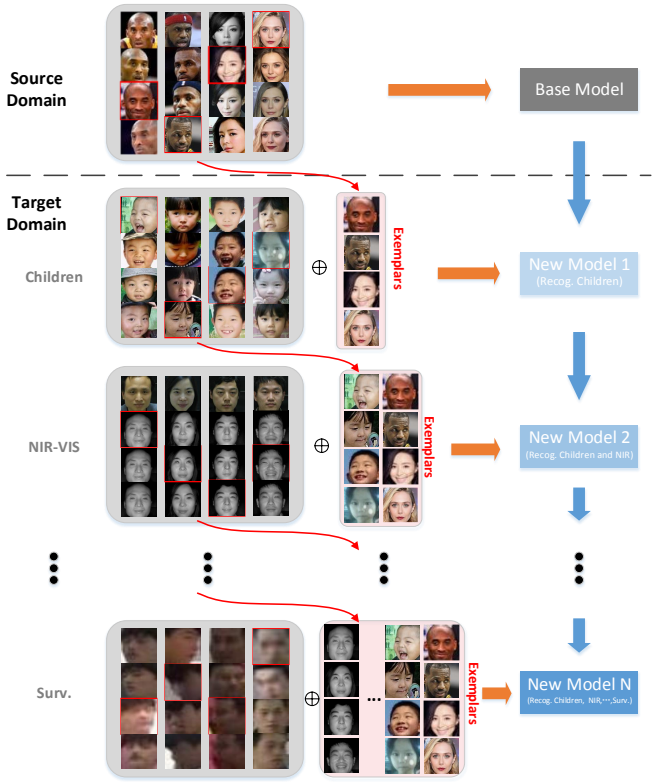


Fig. 1. The process of Single Exemplar Domain Incremental Learning. Starting with a well-trained base model, each time we encounter a new domain, it can adapt to the new domain and preserve the performance of the source domain, and finally, get superior generalization capabilities.

I. Introduction

In recent years, face recognition has made significant progress with the development of deep learning [1]–[3]. However, there remain many challenges in real applications, such as face recognition for the elderly and children, for surveillance scenes and for near-infrared vs. RGB heterogeneous scene, etc. At present, most of the face datasets in academia are celebrity faces collected from the Internet, such as CASIA-WebFace [4], VGGFace2 [5] and MS-Celeb-1M [6]. A well trained model on this data can only achieve good performance in webface scenarios, however, it cannot handle more complicated real-world scenarios, since existing training data cannot cover those unseen domains. A common way to cope with a new domain is to fine-tune the base model with the target-domain data. However, this brings up a problem that fine-tuning on the target domain will seriously degrade the performance on the source domain, which is usually called catastrophic forgetting. A straightforward solution is the joint training strategy which fine-tunes the model on both source and target domains simultaneously. However, it takes a lot of time to fine-tune the model on the combined data since the source data is generally very large. Besides, in the real application, we expect the model can quickly adapt to the target domain during deployment. Even worse, it will take

Hao Liu, Xiangyu Zhu and Zhen Lei are with Center for Biometrics and Security Research (CBSR) & National Laboratory of Pattern Recognition (NLPR), Institute of Automation Chinese Academy of Sciences (CASIA) and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China (e-mail: {hao.liu2016, xiangyu.zhu, zlei}@nlpr.ia.ac.cn). Zhen Lei is the corresponding author.

Dong Cao is with the ByteDance AI Lab, Beijing, China (e-mail: caodong.c@bytedance.com). This work was done when Dong Cao worked at CASIA.

Stan Z. Li is with the Westlake University, Hangzhou, China (e-mail: Stan.ZQ.Li@westlake.edu.cn).

	Transfer Learning (Fine Tuning)	Incremental Learning	Joint Training	Fast Adapting without Forgetting
target domain performance	best	good	best	✓best
source domain performance	X worst	X bad	best	✓best
training efficiency	fast	fast	X slow	✓fast
storage requirement	no	little	X large	✓little

Fig. 2. Fast Adapting without Forgetting is a new transfer learning without forgetting method to fast adapt models to new domains. Compared with existing knowledge transferring methods like transfer learning, incremental learning, and joint learning, our method achieves high performance on both source and target domains with little data storage and computation cost.

huge training time and data storage if there are multiple target domains which need to be adapted.

To reduce the computational cost and data storage when deploying face systems, we introduce a new task: Single Exemplar Domain Incremental Learning (SE-DIL), given one sample per class in source domain and data from target domain, a high-performance base model from source domain is adapted to the target domain, aiming to achieve high performance in target domain and keep the performance in source domain. When multiple target domains are given, we expect that the model can generalize well to them. Figure 1 shows the process of SE-DIL.

In this paper, the SE-DIL task is presented in face recognition as follows: for the source domain, the dataset is constructed by the celebrity faces from the Internet (e.g., MS-Celeb-1M [6]). On the other hand, we select children faces, near infrared heterogeneous faces and surveillance faces as target domains. Note that the aforementioned children face dataset, named KidsFace, is collected by ourselves. KidsFace contains 12,444 identities (10,444 for training and 2,000 for test) and 354K images. We claim that KidsFace is a major contribution to the face recognition society because: (1) most face models work much better on adult faces than child one; (2) children faces are scarce since they cannot be easily collected using search engines by names like celebrity.

To deal with SE-DIL in face recognition, we propose a novel method, Fast Adapting without Forgetting (FAwF), with three components: margin-based exemplar selection, prototype-based class extension and hard&soft knowledge distillation. Figure 2 shows the comparison of our FAwF with other knowledge transferring methods.

Our contributions can be summarized as:

(1) We introduce a new and challenging task Single Exemplar Domain Incremental Learning for a practical application of face recognition, which aims to quickly adapt a high performance base model to the target domain and keep the performance on all previous domains at the same time.

(2) In response to SE-DIL, we propose a method FAwF with three components: margin-based exemplar selection, prototype-based class extension and hard&soft knowledge distillation. Based on that, we can conveniently adapt the model to a new domain and keep the source-domain performance with little data storage and computation cost.

(3) To study this problem in face recognition, we collected a large-scale database of children faces with 12,444 identities, named KidsFace, which is the first large-scale children database to our best knowledge. We will release the entire KidsFace, when this paper is published.

(4) Sufficient experiments on KidsFace, CASIA NIR-VIS 2.0, QMUL-SurvFace for target domain and LFW, CALFW, CPLFW, CFP-FP, AgeDB-30, IJB-C, MegaFace for source domain show that our method not only achieves high performance on the target domain but also keeps the performance on the source domain with only one sample per old class retained.

II. Related Works

Our FAwF approach mainly builds on the insights of three related problems to address Single Exemplar Domain Incremental Learning: face recognition, incremental learning and transfer learning.

A. Face Recognition

Thanks to the development of deep learning, face recognition has made unprecedented progress. DeepFace [7] first introduces CNN model into face recognition. DeepID series [8]–[10] explores multiple network architectures to improve performance. FaceNet [3] tries a mount of network structures to explore the trade-off between computation and accuracy. Center loss [11] proposes to learn the class-specific feature centers to make features more compact in the embedding space. The L2-softmax [12] and NormFace [13] add a L2-constraint on features and weights to promote the under-represented classes. Recently, enhancing cosine and angular margins between different classes is found to be effective in improving feature discrimination. A-Softmax [14] adds multiplicative angular margin to each identity to improve feature discrimination. CosFace [15] and AM-Softmax [16] add additive cosine margin for better optimization. ArcFace [17] moves the additive cosine margin into angular space to get clear geometrical interpretation and better performance on a series of face recognition benchmarks. AdaptiveFace [18] changes the fixed margin to be learnable and class-related, further squeezing the intra-class variations especially for poor classes.

B. Incremental Learning

Incremental learning has been a long standing problem [19]–[26]. Some parameter based methods [27]–[29] try to estimate the importance of each parameter in the base model and try to change those significant parameters as small as possible. However, it is difficult to find a reasonable metric

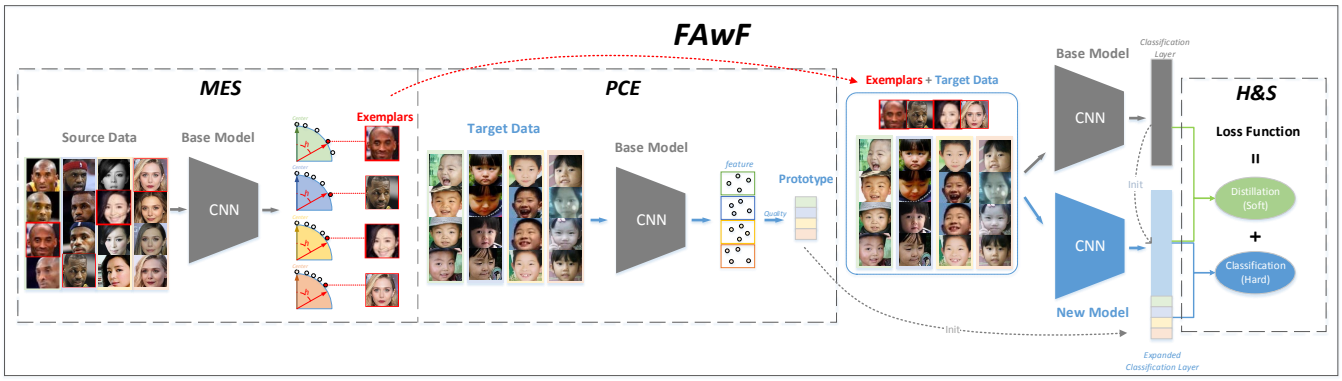


Fig. 3. Overview of our Fast Adapting without Forgetting. It consists of Margin-based Exemplar Selection (MES), Prototype-based Class Extension (PCE) and Hard&Soft Knowledge Distillation (H&S). The base model is not updated during training.

to evaluate all parameters. For class incremental learning, LwF [30] first introduces knowledge distillation [31] to preserve the knowledge of the base model. iCaRL [32] proposes a method to select a small number of exemplars from each old class to preserve old knowledge. EEIL [33] introduces balanced fine-tuning to alleviate the imbalance between old and new classes. BiC [34] adds a bias correction layer to correct the bias between old and new data.

C. Transfer Learning

Transfer learning aims to address the problem when the distribution of the training data from the source domain is different from that of the target domain [35], [35]–[41]. Fine-tuning a pre-trained network model such as ImageNet on a new dataset is the most common strategy for knowledge transfer in the context of deep learning. Most literature in this domain analyzes the effect of pretraining on large-scale datasets with respect to network architectures, network layers, and training tasks [38], [39]. Methods have been proposed to fine-tune all network parameters [42] or only the parameters of the last few layers [35]. [36] investigates several regularization schemes that explicitly promote the similarity of the fine-tuned model with the original pre-trained model.

III. Single Exemplar Domain Incremental Learning

In this section, we first introduce a new task Single Exemplar Domain Incremental Learning in section III-A. In section III-B, we detail our proposed method Fast Adapting without Forgetting to cope with Single Exemplar Domain Incremental Learning.

A. Problem Description

Given a base model $Net(\theta_s)$ trained on the source-domain and the data D_t from the target domain which contains M_t samples, our goal is to adapt the base model to the target domain while keeping the performance on the source domain as high as possible with few examples per class. We denote these exemplars $E_s = \{(x_s^i, y_s^i), x_s^i \in D_s, 1 \leq y_s^i \leq N_s\}$, from the source domain data D_s. x_s^i and y_s^i are the image and label in source domain, respectively. N_s is the number of source domain classes. Note that, D_s and D_t have no class overlap. When there is only a single exemplar per source domain class, we call this task as Single Exemplar Domain Incremental Learning (SE-DIL).

In this paper, we study SE-DIL in the context of face recognition. Storing one sample for each old class can effectively save data storage and training time. Note that conventional incremental learning methods [32]–[34] always need to retain at least 20 samples per old classes.

B. Fast Adapting without Forgetting

In this work, we propose the Fast Adapting without Forgetting (FAwF) method, which is designed to address three key issues in SE-DIL. (1) how to select exemplars when only one exemplar per old class can be retained; (2) how to extend the classification layer when the number of target domain classes increases; (3) what loss function to choose for training. To address these three issues, FAwF proposed margin-based exemplar selection, prototype-based class extension, and hard&soft knowledge distillation. Figure 3 shows the overall framework.

1) Margin-based Exemplar Selection: Since only one sample can be reserved for each old class, it is important to select the most valuable exemplars to preserve source-domain knowledge. In incremental learning, most methods [32]–[34] keep the samples that are as close as possible to their class center. However, in the case of SE-DIL, only one sample can be reserved for each source domain class. We argue that the sample closest to the class center is not the best one to provide enough information to the model in the following adaptation process. For the weights W_s of the classification layer of the source domain model, they are equivalent to the class centers or class prototypes for each class. If we keep the closest sample to the class center of each class from the source domain, it will produce very small loss in the consequent target domain training and thus cannot provide more intra-class information of the source domain.

In order to provide more diverse source domain intra-class information in target domain training to preserve source domain performance as much as possible, we propose the

margin-based exemplar selection scheme to explore this idea. With a given margin h, we select the sample whose distance from the class center is closest to h as the exemplar of this class. Specifically, we first extract features f_s^i for all source domain training samples using the source domain model $Net(\theta_s)$. After that, we calculate the class center of each class and the distance between the feature of each sample and its class center. Finally, given the margin parameter h, the sample that is closest to the distance h from its class center is selected as the exemplar e_j of the class j. Algorithm 1 shows the specific process. In the experimental part, we try different margins and find that if each class retains the sample closest to the center of the class, the performance is far worse than retaining a sample with a large margin, and it is not even as good as random selection.

Algorithm 1: Margin-based Exemplar Selection

Input : $Net(\theta_s)$
$D_s = \{(x_s^i, y_s^i), 1 \leq i \leq M_s, 1 \leq y_s^i \leq N_s\}$
Margin h

Output: Selected exemplars E_s

for *each sample* x_s^i *in* D_s **do**
  Extract the feature f_s^i of x_s^i from $Net(\theta_s)$
end
for $j = 1 \ldots N_s$ **do**
  $c_j = Average(f_s^i), y_s^i = j$
  $distance_s^i = ||f_s^i - c_j||, y_s^i = j$
  $e_j = x_s^{\arg\min(|distance_i - h|)}, y_s^i = j$
end
$E_s = \{e_j, 1 \leq j \leq N_s\}$

To further validate our idea, we analyze the average loss of the exemplars with different margins in the beginning and ending of the training process. As shown in the Figure 4(a), it can be seen that the exemplars from the source domain with a small margin give small losses and cannot be further optimized during training. However, the exemplars with a large margin can give larger initial losses and can be optimized together with target samples during training. We think retaining a relative hard sample as the exemplar can provide some intra-class information of the source domain, which benefits for maintaining the performance of the source domain. Besides, we also analyze the effect to the average loss of the target-domain data when choosing the exemplars with different margins. As shown in the Figure 4(b), the exemplar selection strategy does not affect the target domain performance much, which demonstrates that the choice of different margin exemplars only affects the performance on the source domain. In order to further confirm the idea, we randomly select some samples in the training set of the source domain, and calculate their average loss by models trained with different margin exemplars, as shown in the Figure 4(c). It can be seen that when h increases from 0 to 0.35, exemplars with a larger margin can make the model have smaller loss on the samples in the source domain, that is, achieve better results in the retention of the performance on the source domain. For example, the loss produced by the model when $h = 0.35$ is significantly smaller than that produced by the model with $h = 0$. When h is greater than 0.35, the loss increases slightly, but it is still less than the result of $h = 0$.

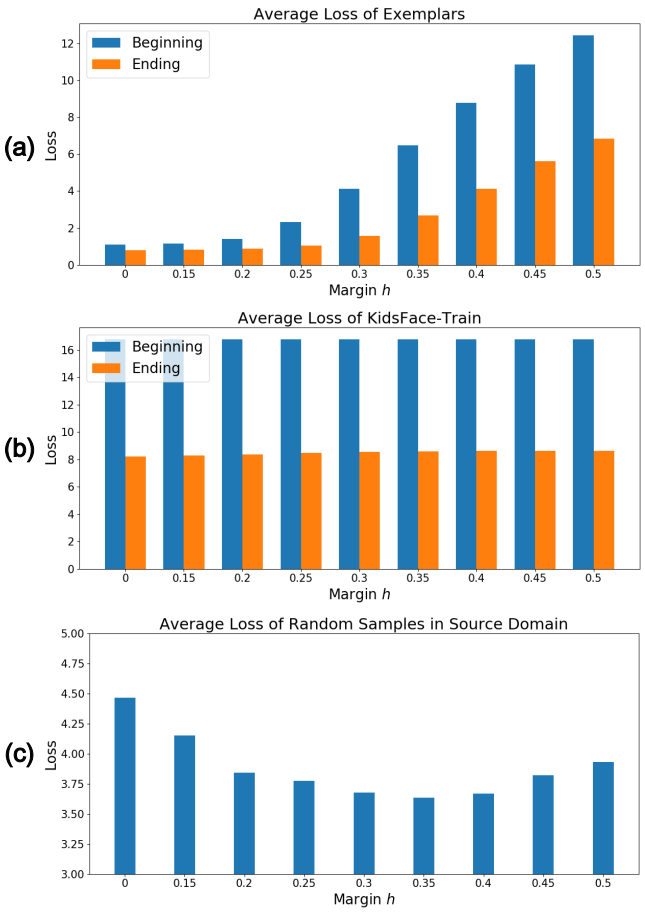


Fig. 4. **(a)** The average loss of the exemplars from the source domain with different margin h in the beginning and ending of the training process. **(b)** The average loss of the samples from the target domain with different margin h in the beginning and ending of the training process. **(c)** The average loss of random samples from the source domain by models trained with different margin h. For source-domain loss, the exemplars with a small margin almost cannot be optimized during training and the exemplars with a large margin can be well optimized. Besides, the exemplar selection strategy does not affect the target domain performance much. Further, the model trained with larger margin exemplars can produce smaller loss for the data in the source domain.

2) Prototype-based Class Extension: Since face recognition is an open-set classification task, in terms of network structure, the classification layer needs to be expanded when continuously adapted to new classes of new domains. For example, if there are N_s classes in the source domain and N_t classes in the target domain, then $N_s + N_t$ classes are needed for the classification layer of the model during adaptation. The previous incremental learning methods [30], [32]–[34] use random initialization for the weights of new classes. However, in the SE-DIL problem, since the previous source domain model already has high performance, in order to preserve more source domain information, we propose Prototype-based Class Extension to initialize the weights of the new classes for the new domain W_t. For this, we introduce the class prototype to initialize the weights of target classes in the classification layer. Specifically, we use the base model $Net(\theta_s)$ to extract

features f_t of all samples of the target domain, and calculate the prototype to represent each new class. The most intuitive way is to average all the features of the class directly, as follows:

$$p_j = \frac{\sum_{y_t^i=j} f_t^i}{N_t}, 1 \le j \le N_t \tag{1}$$

where p_j is the prototype of the target class j and y_t^i is the label of the target samples x_t^i. The W_t is initialized by these prototypes:

$$W_t = [p_1, p_2, \ldots, p_{N_t}] \tag{2}$$

However, the class-center prototype is easily influenced by outliers. In this paper, we introduce the quality factor μ_t^i to weight each image. The quality of each image can be assessed by the feature norm [12] $\mu_t^i = \left\| f_t^i \right\|$. The calculation of the prototypes thus becomes as follows:

$$p_j = \frac{\sum_{y_t^i=j} \mu_t^i f_t^i}{\sum_{y_t^i=j} \mu_t^i}, 1 \le j \le N_t \tag{3}$$

We will discuss the effect of the quality factor in the experiments. Finally, the prototype of the new class is filled into the expanded classification layer. Algorithm 2 summarizes the prototype-based class extension. The proposed extension method not only achieves better performance, but also speeds up the convergence of training.

Algorithm 2: Prototype-based Class Extension

Input : $Net(\theta_s)$
$D_t = \{(x_t^i, y_t^i), 1 \le i \le M_t, 1 \le y_t^i \le N_t\}$
Output: Weight vectors W_t of target domain classes

for *each sample* x_t^i *in* D_t **do**
  Extract the feature f_t^i of x_t^i from $Net(\theta_s)$
end
for $j = 1 \ldots N_t$ **do**
  $p_j = \frac{\sum_{y_t^i=j} \mu_t^i f_t^i}{\sum_{y_t^i=j} \mu_t^i}$, μ_t^i is quality factor
  $w_j = p_j$
end
$W_t = \{w_j, 1 \le j \le N_t\}$

3) Hard&Soft knowledge distillation: In terms of the loss function, we employ two loss functions in our network finetuning. Firstly, the hard label classification loss directly classifies exemplars in D_s and samples in D_t to their corresponding labels. Given an input feature vector f^i with its corresponding label y^i, we use the CosFace [15] for hard label classification:

$$L_c = -\frac{1}{M_t + N_s} \sum_{i=1}^{M_t+N_s} \log \frac{e^{s(\cos(\theta_{iy^i})-m)}}{e^{s(cos(\theta_{iy^i})-m)} + \sum_{j=1,j\neq y^i}^{N_t} e^{s\cos(\theta_{ij})}} \tag{4}$$

$$\cos\theta_{ij} = w_j^{\mathrm{T}} f^i, \|w_j\| = 1, \|f^i\| = 1 \tag{5}$$

where M_t is the number of samples of target domain, N_t is the number of target domain classes, w_j denotes the weight vector of class j, s is the scale factor and m is the margin parameter in CosFace.

Besides the hard-label classification loss, we also propose soft activation distilling loss to better retain the information of the source domain, where the soft activation in the softmax layer on the base model $Net(\theta_s)$ is used to guide the training. Specifically, we denote the output logits of the base model and the new model as $\hat{\mathbf{o}}^{N_s}(x) = [\hat{o}_1(x), ..., \hat{o}_{N_s}(x)]$ and $\mathbf{o}^{N_s+N_t}(x) = [o_1(x), ..., o_{N_s}(x), o_{N_s+1}(x), ..., o_{N_s+N_t}(x)]$, respectively. The distilling loss is formulated as follows:

$$L_d = -\frac{1}{M_t + N_s} \sum_{i=1}^{M_t+N_s} \sum_{j=1}^{N_s} \hat{\pi}_j(x_i) \log[\pi_j(x_i)] \tag{6}$$

$$\hat{\pi}_j(x_i) = \frac{e^{\hat{o}_j(x_i)/T}}{\sum_{k=1}^{N_s} e^{\hat{o}_k(x_i)/T}}, \quad \pi_j(x_i) = \frac{e^{o_j(x_i)/T}}{\sum_{k=1}^{N_s} e^{o_k(x_i)/T}} \tag{7}$$

where T is the temperature scalar. The distilling loss is computed for all samples from the target domain and exemplars from the source domain. Note that the base model is not updated during training.

The overall loss combines the classification loss and the distilling loss as follows:

$$L = L_c + \lambda \cdot L_d \tag{8}$$

where the scalar λ controls the strength of distilling loss L_d, which is discussed in the experiments.

4) Comparison with Previous Methods: For the proposed task SE-DIL in this paper, the most relevant topic is incremental learning, so in this part we compare our FAwF method with the previous incremental learning methods first. LwF [30] does not need to preserve the old class samples and only uses the distillation loss function to preserve the old class information, which leads to its inability to fully preserve the old class performance. In contrast, our FAwF is able to retain high source domain performance with very few source domain samples through a new exemplar selection and classification layer extension methods.

iCaRL [32], EEIL [33] and BiC [34] all need to keep the samples of the old classes to preserve the performance of the old classes in subsequent training, but unlike our approach, they all need to keep a certain number of samples per classes (typically 20), whereas our FAwF only needs to keep single exemplar per class of the source domain. Moreover, about exemplar selection, these methods select the samples closest to the class center as the exemplars. In FAwF, we find that in the case of keeping only one exemplar per class, the sample with a certain margin with the class center can provide more intra-class information of the source domain for subsequent training, thus retaining more source domain performance. In addition, in terms of class extension, these previous methods all use random initialization to extend the weights of new classes, while this paper proposes a new quality factor based prototype initialization method to extend the weights of target domain classes, enabling the model to retain more source domain information. Besides, the task is also related to transfer learning. Compared with transfer learning methods [40], [41], FAwF can transfer to the target domain while preserving the performance of the source domain, and does not require the co-occurrence data to complete the transfer.

TABLE I
TABLE OF CHILD-RELATED FACE DATASETS.

Dataset	No. of Subjects	No. of Images	Avg. Images / Subject	Type	Public
NITL [43]	314	3144	10	child	No
CMBD [44]	106	1060	10	child	No
ITWCC [45]	304	1705	5.6	adult+child	No
CLF [46]	919	3682	4	child	No
FG-NET [47]	82	1003	12.2	adult+child	Yes
CACD [48]	2000	163446	81.7	adult+child	Yes
MORPH-II [49]	13000	55134	4.2	adult+child	Yes
KidsFace	12444	354594	28.5	child	Yes

IV. KIDSFACE DATASET

In this work, we further collect a children face database to validate the effectiveness of proposed method. Since the children data is scare and difficult to collect, the current public webface dataset contains only a very small number of children faces, which in turn leads to the large bias between the domains of webface samples and children face images. This is directly reflected in the poor performance of the well-trained webface model on the children test set, which is shown in the section V-B. We use the children face recognition as the first target domain and examine the performance of proposed method. To this end, we collect and try our best to clean up to get a database with 354K images from 12,444 identities, named KidsFace database in this work.

Table I shows the information for different child-related face databases. It can be seen that the publicly available data sets containing children are all age-related databases (FG-NET [47], CACD [48], Morph-II [49]), most of which are adult images, not specifically for children. There are three children face databases (NITL [43], CMBD [44], CLF [46]), but not publicly available, containing relative limited subjects, with the largest being 919 identities (CLF [46]). In contrast, the collected KidsFace is not only the largest in terms of data size (354k images) compared to previous databases, but also the database specifically for children faces. It will contribute to the future study on children's face recognition and cross-domain and multi-domain face recognition. We will release the entire KidsFace dataset, when this paper is published[1].

Figure 5 shows some samples in KidsFace. Then we selected $2K$ identities of them as the test set, named KidsFace-Test, and the remaining $10,444$ identities as training set, named KidsFace-Train. Within our knowledge, this is the first large-scale children face database for face recognition.

Fig. 5. Some samples in KidsFace. Images on the same line belong to the same identity.

V. EXPERIMENTS

A. Experimental Settings

Preprocessing We detect faces by the FaceBox [50] detector and localize 5 landmarks (two eyes, nose tip and two mouth corners) by a simple 6-layer CNN. All the faces are normalized by similarity transformation and cropped to 112×112 RGB images.

CNN Architecture PyTorch [51] is used to implement our proposed methods. All CNN models in the experiments follow the same architecture in this paper, which is a 50-layer residual network [1] same as LResNet50E-IR in [17]. It has four residual blocks and finally gets a 512-dimensional feature. The networks are trained on TITANX GPUs and the batch size is set to fill all the GPU memory.

Training Data In this paper, for the base model of source domain, we trained it on MS1M-RetinaFace [52] which is a clean version of the MS-Celeb-1M dataset [6]. In total, there are 5.1 million images of $93K$ identities remaining. For the target domain training, we use the KidsFace-Train, CASIA NIR-VIS 2.0 [53] and QMUL-SurvFace [54] separately to learn the target domain information while preserving information about the source domain. CASIA NIR-VIS 2.0 [53] is the mostly used near-infrared heterogeneous face dataset because it is the largest public and most challenging near-infrared database. It is collected in four recording sessions from 2007 to 2010. There are large variations of the same identity, including lighting, expression, pose, and distance. The total number of the subjects in this database is 725. Each subject has 1-22 VIS and 5-50 NIR images. Since each image is randomly gathered, NIR and VIS images have no one-to-one correlation. QMUL-SurvFace [54] is a challenging low-resolution surveillance face dataset, in which low-resolution face images are native and not synthesised by artificial down-

[1]KidsFace will be available soon at http://www.cbsr.ia.ac.cn/users/kidsface/main.htm

TABLE II
THE PERFORMANCE (%) OF BASELINE MODELS ON KIDSFACE-TEST (TAR @ 10^{-5} FAR), LFW, LFW BLUFR (TAR @ 10^{-4} FAR), CALFW, CPLFW, CFP-FP, AGEDB-30, IJB-C (TAR @ 10^{-4} FAR) AND THE TRAINING TIME OF EACH MODEL. 1-ST AND 2-ND BEST RESULTS ARE IN BOLD/UNDERLINE RESPECTIVELY.

Model	Target Domain 1	Source Domain							Training Time(days)
	KidsFace -Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C	
JT(Upper Bound)	90.239	99.75	99.84	95.98	92.88	98.11	98.03	95.19	2.4
BaseS	53.687	99.73	**99.84**	**95.98**	**92.63**	**98.11**	**98.03**	**95.02**	5.5
BaseT	70.872	90.42	25.02	74.36	64.48	68.08	66.20	35.48	0.3
FT	84.661	96.08	60.62	84.43	75.53	79.47	79.40	78.84	0.16
FAwF	**86.846**	**99.75**	99.79	95.95	92.19	97.82	97.91	94.37	**0.1**

sampling of native high-resolution images. It contains 463,507 face images of 15,573 distinct identities captured in real-world uncooperative surveillance scenes across wide space and time. Besides, during training, these face images are horizontally flipped randomly for data augmentation.

Evaluation Setup For each image, we extract features only from the original image as the final representation. The score is measured by the cosine distance of two features. Finally, face verification and identification are conducted by thresholding and ranking the scores. We use KidsFace-Test, CASIA NIR-VIS 2.0 [53], QMUL-SurvFace [54] as three sequential target domains and use LFW [55], LFW BLUFR [56], CALFW [57], CPLFW [58], CFP-FP [59], AgeDB-30 [60], IJB-C [61] and MegaFace [62] (clean version from ArcFace [17]) as the databases from source domain.

B. Baseline Models

Base model in source domain For the base model of the source domain(BaseS), we use CosFace [15] to train the model on MS1M-RetinaFace from scratch. The m of CosFace is 0.4 and s is 64. It is the staring model that used for fine-tuning on the target domain in all the following experiments.

Base model in target domain For the sake of comparison, we also used KidsFace-Train to train a model from scratch with the same loss function CosFace, named BaseT.

Joint training For joint training (JT), we use all the data from MS1M-RetinaFace and KidsFace-Train to fine-tune the BaseS. For the extension of the classification layer, we use BaseS to get the prototypes to fill the weight vectors of the classification layer. In this way, we simulate the process of joint training when encountering a new domain and having the base model BaseS in the real application. We treat the results of joint training as the performance **upper bound** in this paper.

Fine-tuning For fine-tuning model (FT), we use KidsFace-Train to fine-tune the base model BaseS without initializing the weight vectors of KidsFace-Train classes.

We evaluate these benchmark models on both the KidsFace-Test of the target domain and the LFW [55], CALFW [57], CPLFW [58], CFP-FP [59], AgeDB-30 [60], IJB-C [61] of the source domain. The results are shown in Table II. It can be seen that the BaseS has high performance on the source domain but poor performance on the target domain. The performance of the BaseT is improved compared to the BaseS in target domain, but the performance on the source domain is extremely poor. Although JT achieves the best performance on both the target domain and the source domain, it utilizes all the source domain data and suffers from large data storage and training time. The performance of FT in the target domain is relatively high, but the performance on the source domain is much lower than that of BaseS. Our proposed FAwF achieves a significant improvement on the target domain compared to BaseS, while the performance on the source domain is almost not degraded compared to BaseS (the decline is less than 0.8% on all source domain protocol). Compared with JT, FAwF has a significant reduction in training time.

Method	IRIS	
	1E-1	1E-2
Contrastive	52.62	32.88
Triplet	47.06	29.64
LwF [30]	50.66	23.75
iCaRL [32]	53.38	25.48
EEIL [33]	55.50	27.92
BiC [34]	58.21	33.50
FAwF	**58.48**	**33.66**

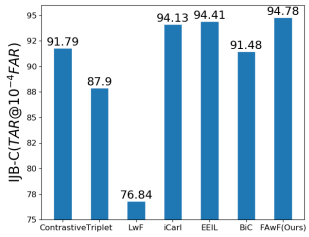


Fig. 6. The first-stage performance of our FAwF and other comparison methods on IRIS (target domain) and IJB-C (source domain).

C. Overall Benchmark Comparisons

In this part, we evaluate the performance of proposed FAwF on sequential domains, i.e., from webface (source domain) to children faces (1st target domain), and to NIR-VIS heterogeneous faces (2nd target domain), and to surveillance faces (3rd target domain). Besides, we use a large number of generic face test sets to evaluate the performance of our method on the source domain. The base model of all the methods in the first stage is BaseS, and the base models in the second and third stage are the models obtained by each method after the previous stage training. We compare the methods of metric learning and incremental learning with our FAwF. For metric learning, we use contrastive loss and triplet loss to fine-tune the BaseS with KidsFace-Train, respectively. For incremental learning methods (iCaRL [32], EEIL [33] and BiC [34]), we reserve one sample per old class and retain the sample closest

TABLE III
THE FIRST-STAGE PERFORMANCE OF OUR FAWF AND OTHER COMPARISON METHODS ON KIDSFACE-TEST (TAR @ 10^{-5} FAR), LFW, LFW BLUFR (TAR @ 10^{-4} FAR), CALFW, CPLFW, CFP-FP, AGEDB-30, IJB-C (TAR @ 10^{-4} FAR) AND AND MEGAFACE. "E" INDICATES FINE-TUNING WITH EXEMPLARS OF SOURCE DOMAIN.

Method	Target Domain 1	Source Domain								
	KidsFace-Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C	MF1 Rank 1	MF1 Veri.
Contrastive	83.753	99.22	95.53	93.15	87.75	94.70	94.06	83.02	83.815	85.484
Trplet	85.057	99.20	94.22	92.96	86.25	93.48	92.18	81.80	85.823	86.676
LwF [30]	84.977	95.83	52.80	84.63	73.96	77.15	79.71	59.04	40.145	37.355
iCaRL [32]	73.894	99.68	99.35	95.76	90.38	95.97	96.95	91.74	94.117	95.215
EEIL [33]	73.585	99.70	99.49	95.70	90.70	95.57	97.31	91.91	95.162	96.030
BiC [34]	81.722	99.45	98.72	95.33	87.65	91.54	95.95	79.69	90.681	92.508
FAwF	**86.846**	**99.75**	**99.79**	**95.95**	**92.19**	**97.82**	**97.91**	**94.37**	**96.899**	**97.123**

TABLE IV
THE SECOND-STAGE PERFORMANCE OF OUR FAWF AND OTHER COMPARISON METHODS ON CASIA NIR-VIS 2.0 (TAR @ 10^{-5} FAR), KIDSFACE-TEST (TAR @ 10^{-5} FAR), LFW, LFW BLUFR (TAR @ 10^{-4} FAR), CALFW, CPLFW, CFP-FP, AGEDB-30 AND IJB-C (TAR @ 10^{-4} FAR). "E" INDICATES FINE-TUNING WITH EXEMPLARS OF SOURCE DOMAIN AND THE FIRST TARGET DOMAIN.

Method	Target Domain 2	Target Domain 1	Source Domain						
	CASIA NIR-VIS 2.0	KidsFace-Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C
Contrastive	99.291	30.585	98.72	87.70	90.96	85.21	92.18	91.35	78.01
Trplet	99.267	57.864	98.93	92.48	92.11	84.23	90.72	91.01	71.57
LwF [30]	97.881	70.706	96.08	54.91	84.30	73.16	75.87	78.91	58.10
iCaRL [32]	95.586	72.425	99.68	99.52	95.61	90.63	95.84	97.18	91.20
EEIL [33]	93.862	70.412	99.72	99.45	95.68	90.20	95.41	97.21	89.83
BiC [34]	94.660	72.596	99.50	98.53	95.05	86.80	90.01	95.65	76.62
FAwF	**99.629**	**84.731**	**99.73**	**99.78**	**95.95**	**92.36**	**98.05**	**97.96**	**94.33**

TABLE V
THE THIRD-STAGE PERFORMANCE OF OUR FAWF AND OTHER COMPARISON METHODS ON QMUL-SURVFACE (TAR @ 10^{-2} FAR), CASIA NIR-VIS 2.0 (TAR @ 10^{-5} FAR), KIDSFACE-TEST (TAR @ 10^{-5} FAR), LFW, LFW BLUFR (TAR @ 10^{-4} FAR), CALFW, CPLFW, CFP-FP, AGEDB-30 AND IJB-C (TAR @ 10^{-4} FAR). "E" INDICATES FINE-TUNING WITH EXEMPLARS OF SOURCE DOMAIN AND THE FIRST TARGET DOMAIN.

Method	Target Domain 3	Target Domain 2	Target Domain 1	Source Domain						
	QMUL-SurvFace	CASIA NIR -VIS 2.0	KidsFace -Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C
Contrastive	44.2	1.216	10.974	86.07	4.78	64.81	63.48	65.00	62.01	14.62
Trplet	50.0	1.006	17.139	91.32	20.24	70.93	69.10	72.60	64.43	24.22
LwF [30]	54.7	0.483	14.441	78.72	2.77	57.98	60.38	61.94	51.28	7.72
iCaRL [32]	56.5	54.977	63.519	99.60	99.29	95.30	87.21	88.48	96.36	86.12
EEIL [33]	57.5	54.131	64.492	99.63	99.27	95.53	87.05	88.50	96.56	86.30
BiC [34]	45.5	65.568	61.374	99.08	96.62	93.93	80.60	78.02	93.48	74.73
FAwF	**59.3**	**97.906**	**71.004**	**99.70**	**99.74**	**95.46**	**91.28**	**96.83**	**97.15**	**90.43**

to the class center in their way, i.e. $h = 0$ in margin-based exemplar selection. For LwF [30], no exemplar is needed.

Table III showcases the results of these methods after the first stage. Compared with other methods, the proposed FAwF not only achieves the best performance on the target domain test set (86.846% on KidsFace-Test), but also achieves the best performance on all source domain benchmarks (*e.g.*, 96.899% on MF1 Rank1). The methods of metric learning achieve good performance on the target domain, but the performance on the source domain drops significantly, reflecting the catastrophic forgetting phenomenon. For incremental learning methods, most of them can maintain high source domain performance, but the performance on the target domain is not satisfactory. For example, EEIL can reach 95.162% on MF1 Rank1, but only 73.585% on the target domain.

In addition, we conduct the first stage experiments on a

thermal face dataset, IRIS Face Database [63]. Since there is no official test protocol for this database and no distinction between training and test sets, we have designed a test protocol for this database. Specifically, we used the data of the first 21 identities ("Balage" to "priya") in the database as the training set and the data of the remaining 10 identities ("Rangan" to "vivek") as the test set. Figure 6 shows the results of different methods on this database. Our FAwF has the highest performance in both source and target domains compared to other methods, which is consistent with the results on KidsFace.

Table IV shows the results of these methods after continuing the second stage near-infrared heterogeneous face training based on the first stage. It can be seen that our FAwF still obtains the highest performance on the second target domain, and also achieves the best performance on the first target domain and source domain. The metric learning methods can still achieve good performance on the target domain at this stage, but the performance on the source domain and the previous target domain is significantly reduced. For the incremental learning methods, they still cannot achieve satisfactory performance on the target domain.

Table V shows the third stage results. The proposed FAwF can still achieve the highest performance in all target domains and source domains in the third stage. Unlike the previous two stages, the incremental learning methods also achieve relatively good performance on both the source and target domains.

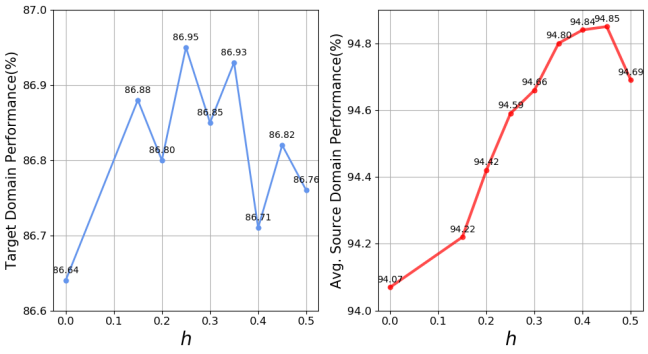


Fig. 7. Target domain performance (%) of FAwF with different h on KidsFace-Test (TAR @ 10^{-5} FAR) and average performance on three source domain protocols (CPLFW, CFP-FP and IJB-C (TAR @ 10^{-4} FAR)).

D. Exploratory Experiments

Effect of h in Margin-based Exemplar Selection. Since only one sample can be reserved for each old class of the source domain, it is essential to decide which sample is selected as the exemplar for that class. In this section, we build an experiment to explore the effects of margin-based exemplar selection with different margin h. By varying h from 0 to 0.5, we fine-tune the base model BaseS on the KidsFace-Train using our proposed FAwF. In this part, we only use CosFace as the loss function. We evaluate the performance on the target domain on KidsFace-Test and the performance on the source domain on CPLFW, CFP-FP and IJB-C. The

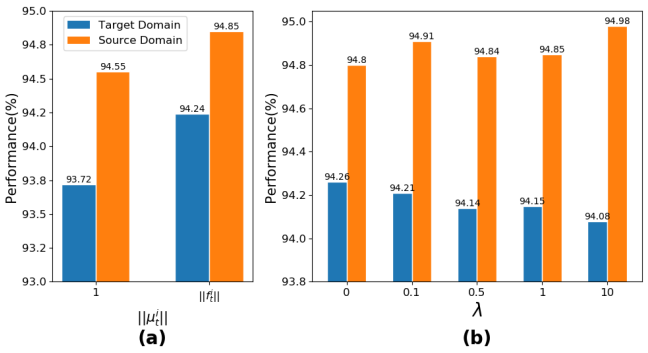


Fig. 8. (a) Target domain performance (%) of FAwF with different μ_t^i on KidsFace-Test (TAR @ 10^{-5} FAR) and average performance on three source domain protocols (CPLFW, CFP-FP and IJB-C (TAR @ 10^{-4} FAR)). (b) Target domain performance (%) of FAwF with different λ on KidsFace-Test (TAR under 10^{-4} FAR) and average performance on three source domain protocols.

results are shown in Figure 7. We can see that as h increases, all models can achieve high performance (over 86.5%) on the KidsFace-Test. As for the performance of the source domain, we observe a obvious trend that the performance gradually increases with the increasing of h and get saturated when $h = 0.45$, then slightly decreasing. The common used $h = 0$, which is the class center, is not a good choice in this task. We also add a control experiment with randomly selected samples as exemplars. The source domain performance result is that retaining the sample closest to the center of the class (94.07%) is not even as good as retaining a random sample per old class (94.59%). This validates our hypothesis that hard exemplar can bring more information about source domain compared with simple exemplar during fast adapting.

Effect of μ_t^i in Prototype-based Class Extension. To investigate the effect of adding the quality factor μ_t^i, we compared the performance without using μ_t^i ($\mu_t^i = 1$) and using the norm of feature as the quality factor ($\mu_t^i = \left\|f_t^i\right\|$), as shown in the Figure 8(a), where h is set to 0.35. It can be seen that the performance of using the quality factor has a steady improvement on both the target domain and the source domain.

Effect of λ in H&S Knowledge Distillation. In order to explore the effect of distilling loss, we varied the λ from 0 to 10 to fine-tune the model with FAwF. h is set to 0.35. Figure 8(b) shows the results for different lambdas. When $\lambda = 0$, the distill loss is not used. As we can see, the performance of the target domain decreases slightly with λ increasing, while the performance of the source domain is slightly improved. Overall, the impact of λ on performance is small. In other experiments when using hard&soft knowledge distillation, we set the λ to 1.

E. Ablation Study

To demonstrate the effectiveness of the three components in our framework, we run a number of ablations to analyze the improvements from margin-based exemplar selection, prototype-based class extension and hard&soft knowledge distillation, respectively. From Table VI, we can see that

TABLE VI
ABLATION STUDY ON KIDSFACE-TEST (TAR @ 10^{-5} FAR), LFW, LFW BLUFR (TAR @ 10^{-4} FAR), CALFW, CPLFW, CFP-FP, AGEDB-30, IJB-C (TAR @ 10^{-4} FAR) AND AND MEGAFACE. PCE INDICATES THE PROTOTYPE-BASED CLASS EXTENSION, MES INDICATES THE MARGIN-BASED EXEMPLAR SELECTION, H&S INDICATES THE HARD&SOFT KNOWLEDGE DISTILLATION. H INDICATES THAT ONLY THE CLASSIFICATION LOSS FUNCTION IS USED.

PCE	MES	H&S	Target Domain 1	Source Domain								
			KidsFace-Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C	MF1 Rank 1	MF1 Veri.
-	-	H	86.039	98.85	86.32	94.21	85.30	87.87	93.70	85.88	80.466	82.978
✓	-	H	86.645	99.68	99.66	95.86	91.61	96.97	97.23	93.63	94.598	95.350
✓	✓	H	86.825	99.73	99.78	**95.96**	**92.53**	97.65	97.66	94.36	96.645	96.682
-	-	H&S	73.641	99.70	99.41	95.56	90.49	95.34	97.24	91.04	92.688	92.893
✓	-	H&S	**87.045**	99.73	99.69	95.76	91.76	97.45	97.28	93.93	95.294	96.385
✓	✓	H&S	86.846	**99.75**	**99.79**	95.95	92.19	**97.82**	**97.91**	**94.37**	**96.899**	**97.123**

improvement from prototype-based class extension is the most obvious on both target and source domains (from 86.039% to 86.645% in KidsFace-Test and more than 10 percentage point increase in LFW BLUFR and MF1). Besides, margin-based exemplar selection has a significant improvement on the performance of the source domain (from 94.598% to 96.645% in MF1 Rank 1). Finally, Using hard&soft knowledge distillation alone can significantly improve the performance of the source domain (from 85.88% to 91.04% in IJB-C), but it will harm the performance of the target domain (from 86.039% to 73.641% in KidsFace-Test). When combining H&S with the other two parts, it can still bring a slight improvement for source domain performance. When the three parts are combined, FAwF can greatly improve the performance of the source domain while ensuring that the performance in the target domain is higher than 86.8%.

VI. CONCLUSION

In this paper, we introduce a new challenging task Single Exemplar Domain Incremental Learning (SE-DIL), which utilizes target-domain data and a little source-domain data to quickly improve the performance on the target domain based on a well-trained base model while keeping the performance on the source domain. To cope with SE-DIL, we propose the Fast Adapting without Forgetting (FAwF) method, which consists of margin-based exemplar selection, prototype-based class extension and hard&soft knowledge distillation. Besides, for studying the SE-DIL in face recognition, we collected a large-scale children face dataset KidsFace with 12K identities. Extensive experiments show that proposed FAwF can well maintain the source domain performance with only one sample per source domain class and outperforms the state-of-the-art methods on both target and source domains.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects #61876178, #61872367, #61806196, #61976229. The authors also thank Dr. Guosheng Hu to help improve the quality of this paper.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *computer vision and pattern recognition*, pp. 770–778, 2016.

[2] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[4] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[6] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[8] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy *et al.*, "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2403–2412.

[9] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[10] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[12] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[13] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: l 2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1041–1049.

[14] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[15] H. Wang, Y. Wang, Z. Zhou, X. Ji, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

[16] F. Wang, W. Liu, H. Liu, and J. Cheng, "Additive margin softmax for face verification," *Signal Processing Letters, IEEE*, vol. 25, pp. 926–930, 2018.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[18] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "Adaptiveface: Adaptive margin and sampling for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 947–11 956.

[19] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Advances in neural information processing systems*, 2001, pp. 409–415.

[20] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, vol. 31, no. 4, pp. 497–508, 2001.

[21] D. S. Tan, Y.-X. Lin, and K.-L. Hua, "Incremental learning of multi-domain image-to-image translations," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[22] L. Yu, B. Twardowski, X. Liu, L. Herranz, K. Wang, Y. Cheng, S. Jui, and J. v. d. Weijer, "Semantic drift compensation for class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6982–6991.

[23] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 208–13 217.

[24] Y. Liu, Y. Su, A.-A. Liu, B. Schiele, and Q. Sun, "Mnemonics training: Multi-class incremental learning without forgetting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 245–12 254.

[25] J. He, R. Mao, Z. Shao, and F. Zhu, "Incremental learning in online scenario," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 926–13 935.

[26] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 183–12 192.

[27] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[28] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3987–3995.

[29] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[30] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[32] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[33] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 233–248.

[34] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 374–382.

[35] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[36] L. Xuhong, Y. Grandvalet, and F. Davoine, "Explicit inductive bias for transfer learning with convolutional networks," in *International Conference on Machine Learning*, 2018, pp. 2825–2834.

[37] M. Meng, M. Lan, J. Yu, and J. Wu, "Coupled knowledge transfer for visual data recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[39] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3712–3722.

[40] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4376–4386, 2019.

[41] J. Lin, L. Zhao, Q. Wang, R. Ward, and Z. J. Wang, "Dt-let: Deep transfer learning by exploring where to transfer," *Neurocomputing*, 2020.

[42] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[43] L. Best-Rowden, Y. Hoole, and A. Jain, "Automatic face recognition of newborns, infants, and toddlers: A longitudinal evaluation," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–8.

[44] P. Basak, S. De, M. Agarwal, A. Malhotra, M. Vatsa, and R. Singh, "Multimodal biometric recognition for toddlers and pre-school children," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 627–633.

[45] K. Ricanek, S. Bhardwaj, and M. Sodomsky, "A review of face recognition against longitudinal child faces," *BIOSIG 2015*, 2015.

[46] D. Deb, N. Nain, and A. K. Jain, "Longitudinal study of child face recognition," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 225–232.

[47] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on pattern Analysis and machine Intelligence*, vol. 24, no. 4, pp. 442–455, 2002.

[48] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *European conference on computer vision*. Springer, 2014, pp. 768–783.

[49] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 341–345.

[50] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: A cpu real-time face detector with high accuracy," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 1–9.

[51] "Pytorch," https://pytorch.org/.

[52] "Ms1m-retinaface," https://github.com/deepinsight/insightface/tree/master/iccv19-challenge.

[53] S. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 348–353.

[54] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *arXiv preprint arXiv:1804.09691*, 2018.

[55] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[56] S. Liao, Z. Lei, D. Yi, and S. Z. Li, "A benchmark study of large-scale unconstrained face recognition," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–8.

[57] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[58] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, pp. 18–01, 2018.

[59] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

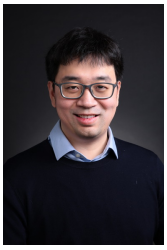
[60] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–59.

[61] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa janus benchmark-b face dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 90–98.

[62] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[63] R. Miezianko, "Ieee otcbvs ws series bench," *Terravic research infrared database*, vol. 2, 2006.

Hao Liu Hao Liu received the B.S. degree in Beijing Institute of Technology (BIT) in 2016. Since September 2016, he has been a Ph.D. candidate at the National Laboratory of Pattern Recognition (NLPR), Instituteof Automation, Chinese Academy of Science (CASIA). His research interests include computer vision, pattern recognition, especially with a focus on face recognition.



Xiangyu Zhu Xiangyu Zhu received the BS degree in Sichuan University (SCU) in 2012, and the PhD degree from Institute of Automation, Chinese Academy of Sciences, in 2017, where he is currently an associate professor. His research interests include pattern recognition and computer vision, in particular, 3D reconstruction, 3D face model and face recognition.

Zhen Lei Zhen Lei received the BS degree in automation from the University of Science and Technology of China (USTC), in 2005, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently a professor. He has published more than 160 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an area chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015. He is the winner of 2019 IAPR YOUNG BIOMETRICS INVESTIGATOR AWARD. He is a senior member of the IEEE.

Dong Cao Dong Cao received the B.S. degree in Nanjing University of Aeronautics and Astronautics in 2010, the M.S. degree in Hohai University in 2013, and the Ph.D. degree in Chinese Academy of Sciences in 2016. He is now a Senior R&D Engineer of computer vision team in ByteDance, Inc. His research interests are in computer vision, pattern recognition, deep learning, and in particular, face recognition.

Stan Z. Li received the B.Eng degree from Hunan University, the M.Eng degree from National University of Defense Technology, and the Ph.D. degree from Surrey University. He is currently a chair professor in WestLake University. He was a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA) from 2004 to 2018. He was with MSRA as a researcher from 2000 to 2004. Prior to that, he was an associate professor in the Nanyang Technological University. His research interests include image and vision processing, pedestrian recognition and biometrics. He has published more than 300 papers in international journals and conferences, and authored and edited 8 books. He was an associate editor of the IEEE TPAMI and is acting as the editor-in-chief for the Encyclopedia of Biometrics. He served as a program co-chair for ICB 2007, 2009, 2013, 2014, 2015, 2016 and 2018, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to IEEE fellow for his contributions to the fields of face recognition, pattern recognition and computer vision and he is a member of the IEEE Computer Society.